TD : Prise en main d'UNIX et outils d'analyse de séquence sous UNIX

Daniel Gautheret

2009 V.1

Notes :

- Le compte « prof » fait référence à : `~gauthere/TPUNIX`
- Le fichier .bashrc doit contenir :
  `export PATH=.:~gauthere/bin/:$PATH`

1. Prise en main

- Connectez-vous sur votre station Unix avec votre nom et votre mot de passe (communiqués en TD). S'il s'agit de votre première connexion,
- Les commandes Unix sont entrées dans une fenêtre de commande. Sachez créer et supprimer les fenêtres de commande. Voyez bien la différence entre "réduction" et "suppression" des fenêtres.
- Voyez comment rappeler la commande précédente avec la touche "flêche haut", puis "flêche gauche" pour corriger.

2. Fichiers et Répertoires

- Créez un Répertoire `TPUNIX` à la racine de votre Répertoire personnel (`mkdir`)
- Placez-vous dans le répertoire `TPUNIX` (`cd`)
- Vérifiez que vous êtes bien dans ce répertoire (`pwd`).
- Copiez dans `TPUNIX` tous les fichiers se trouvant dans le répertoire `prof` (commande cp).
- Afficher le contenu de ce répertoire. Quelle est la taille des fichiers? Le propriétaire? La date de création? Quels sont les fichiers executables? (commande ls et ls -l).
- Le fichier "`test.pl`" est un programme. Exécutez-le.
- Tout en restant dans le répertoire courant, exécutez la commande "`test2.pl`" qui se trouve dans `"prof"`
- Affichez le contenu du fichier « `16s.seq` » à partir de votre répertoire racine.
- A l'aide de la commande `cat` et des redirections (`>` et `>>`), concatenez le fichier `'16s.seq'` et le fichier 'petiteseq.fasta' dans le fichier `'allseq'`. Affichez le résultat.

3. Edition

- Lancez l'editeur nedit (commande `nedit <nom de fichier> &`). Vérifier toutes les fonctionnalités de base : Entrée de texte, sauvegarde, recherche, couper et coller via les menus et/ou touches contrôle-xxx.
- Voyez comment effectuer les copier/coller d'une fenêtre à l'autre, à l'aide de la souris. *Les fonctions copier/coller sous Unix-Xwindows sont réalisées à l'aide des boutons de la souris. Bouton de gauche pour copier, bouton central (ou deux boutons en même temps) pour coller.*

4. Commandes diverses

- Vous devez être familier avec les commandes suivantes:
  - man (utilisez-le pour comprendre les commandes ci-dessous)
  - ps (quelle option pour voir tous les process?)

- rm (effacez un fichier)
- du (quel espace disque occupez-vous?)
- kill (lancez nedit, voyez son numéro de process avec ps, tuez-le a partir d'une autre fenêtre).

5. Analyse d'un fichier Genbank avec egrep

Objectif: Extraire rapidement les informations présentes dans un fichier Genbank.

- Regardez le fichier "`mgen.gbk`" à l'aide de la commande "`more`". Repérez les séquences protéiques. Que veut dire "CDS"? Que veut dire "complement" après CDS? Où se trouve la séquence nucléotidique? Quelles informations sont disponibles sur chaque gène?
- Avec la commande egrep, comptez les élements suivants
  - les gènes protéiques annotés
  - les gènes protéiques présents sur le brin inverse
  - les tRNA
  - Les gènes prédits par similitude de séquence
  - Les gènes prédits par GeneMark.

  attention aux majuscules/minuscules: utiliser `egrep -i` en cas de doute

6. Récupération de génomes complets via un serveur ftp

La plupart des génomes complètement séquencés sont déposés d'une part dans Genbank et EMBL/EBI, et d'autre part sur les serveurs Web ou ftp des différentes institutions ayant généré ces séquences. Nous allons récupérer un génome sur le site EMBL à l'EBI.

- Lancez votre navigateur et connectez vous sur:
  `ftp://ftp.ncbi.nlm.nih.gov/`

  Notez que le protocole de communication est ici <u>Ftp</u> et non pas http. Nous ne sommes pas sur page Web, mais sur un serveur ftp, conçu pour le transfert de fichiers uniquement. Les liens correspondent à des répertoires sur le disque du serveur ftp

- Ce serveur ftp comporte tous les fichiers de Genbank.. Explorez le répertoire "genomes" qui contient les génomes complets ou en cours de séquençage complet. Identifiez les fichiers au format fasta et embl/gbk.
  - Les fichiers en .Z ou .gz sont des fichiers compressés. Après les avoir téléchargé, il est nécessaire de les décompresser à l'aide du programme uncompress (pour les fichiers .Z) ou gunzip (pour les fichiers .gz).
- Identifiez le répertoire contenant le génome de la bactérie *Haemophilus influenza*.
- Récupérez le fichiers .faa, .fna et .gbk pour cet organisme. Décompressez-les si nécessaire. Que contiennent-ils ?

7. Blast, localement

Blast ne travaille pas sur un fichier au format Fasta. Il faut pré-traiter les fichiers Fasta

avec le programme `formatdb`.

A la racine de votre répertoire, créez un fichier .ncbirc contenant:

```
[NCBI]
   Data=/usr/local/biotools/bin

[BLAST]
BLASTDB=.
```

Ce fichier indique où se trouvent les divers fichiers indispensables à Blast, p. ex. les matrices de substitution

- A l'aide de la commande `formatdb` préparez le fichier Fasta du génome au traitement par Blast. Attention: arguments différents pour les séquences nucléiques et protéiques.
  - o `formatdb –i <protein database> -p T`
  - o `formatdb –i <DNA database> -p F`
- Regardez quels fichiers ont été créés.
- Tapez `blastall` sans argument pour obtenir la liste des arguments de Blast. L'argument -p qui spécifie la version de Blast est indispensable (p. ex`: -p blastp` pour Blast protéine), ainsi que les arguments `-d` et `-i`.
- Avec Blast, recherchez dans le génome de *H. influenzae* des séquences similaires à `16s.seq`.
- Exécutez Blast de façon à n'obtenir que les solutions de E-value inférieure à 10e-4, en redirigeant la sortie vers un fichier. Gardez ce fichier.

8. Alignement de séquences avec Muscle

Objectif: réaliser un alignement en mode local.

- A partir de la séquence abc.fasta (un ABC transporteur de E. coli), identifiez par Blast les ABC transporteurs homologues dans le génome de *H. influenzae*. Récupérez une dizaine d'homologues et sauvegardez-les dans un fichier au format Fasta.
- Lancez Muscle (`muscle`) et alignez les séquences d'ABC transporteurs. Pour le mode d'emploi du programme, tapez `muscle` sans aucun argument.
- Visualisez le fichier d'alignement avec `more`.